УДК 004.934.5

Предсказание акустических характеристик звука для систем TTS

Р. Н. Макаров
(Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к. ф.-м. н., доцент, С. В. Рыбин
(Университет ИТМО, г. Санкт-Петербург)

Введение

Исследования выполнены за счет стартового финансирования университета ИТМО в рамках НИР № 618278 «Синтез эмоциональной речи на основе генеративных состязательных сетей». Целью магистерской диссертации является определение тонального контура методами машинного обучения.

Базовые положения исследования

Ранее, на предыдущих конференциях было рассказано о построении классификаторов по предсказанию грамматических характеристик текста и просодических характеристик звука. Так же в рамках НИР и магистерской диссертации была проделана работа по предсказанию частоты основного тона по алофонной разметке с использованием ансамбльных алгоритмов машинного обучения [1].

На данном этапе была проведена работа по объединению предыдущих частей. Базовым предприятием – Центром Речевых Технологий (ЦРТ) была расширена предыдущая база звуковых данных с соответствующей аллофонной разметкой, с 500 до 1800 файлов. Так же была добавлена эмоциональная составляющая — база разделена на 3 части по соответствующим эмоциональным состояниям диктора (happy, depress и neutral).

Классификаторы, предсказывающие грамматические (POS-tagging) и просодические (предсказание тонального контура) характеристики были улучшены с помощью добавления векторного представления слов — word2vec [2] (ранее был использован обучаемый embedding-слой нейронной сети). Качество классификаторов было улучшено, так как в векторном пространстве word2vec слова располагаются по признаку семантической близости.

Далее с помощью этих классификаторов были добавлены соответствующие признаки к базе данных алофонной разметки. Были построены первые классификаторы на основе LSTM-слоев рекуррентной нейронной сети [3]. С помощью построенных классификаторов предсказываются акустические характеристики звука — частота основного тона, характеристики мел-спектра и энергия сигнала. Однако для получения хорошей точности предсказания акустических характеристик звука необходимо дальнейшее расширение базы.

Выводы

В рамках проведенной работы:

1) С помощью векторного представления слов была улучшена точность предсказания грамматических и просодических признаков;
2) Было произведено расширение текущей базы аудиофайлов с аллофонной разметкой;
3) К основным признакам для предсказания акустических характеристик были добавлены грамматические и просодические признаки. Был построен классификатор на основе LSTM-слоев рекуррентной нейронной сети.

Литература

1. Friedman J. H. Greedy function approximation: a gradioent boosting machine //Annals of statistics. – 2001. – С. 1189-1232.

2. Mikolov T. Et al. Distributed representations of words and phrases and their compositionaly //Advances in neural information processing systems. – 2013. – С. 3111-3119.

3. Hochreiter S., Schmidhuber J. Long short-term memory //Neural computation. – 1997. – Т. 9. – №. 8. – С. 1735-1780.